共変量のある判別問題について (多変量統計解析 II)

山口, 光代

数理解析研究所講究録
247巻 1975年 63-72

共変量のある判別問題について

大阪大 基礎工 山口光代

共変量をもつ2つの正規母集団 $\Pi_i : N_{p+q}\left[\begin{bmatrix}\mu_i \\ \eta\end{bmatrix}, \Sigma\right]$ $i=1,2$ からの N_i 個の観測値 $\begin{bmatrix}x_{i1} \\ y_{i1}\end{bmatrix}, \cdots, \begin{bmatrix}x_{iN_i} \\ y_{iN_i}\end{bmatrix}$ に基づいて、他の観測値 $\begin{bmatrix}x \\ y\end{bmatrix}$ が、Π_1 又は Π_2 のいずれに属するかを判別する問題を考える。ここに

$\mu_i\ (p\times 1)$, $i=1,2$; $\eta\ (q\times 1)$, $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$ $(\overline{p+q} \times \overline{p+q})$ は未知パラメーターで、$\Delta^2 = (\mu_1-\mu_2)'\Sigma_{11.2}^{-1}(\mu_1-\mu_2) > 0$ とする。$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. この判別問題に対して、次の3つの判別統計量を扱う。

[1]. Cochran and Bliss (AMS 1948) によって与えられた線形判別統計量

$$W^* = \left[x^* - \frac{1}{2}(\bar{x}_1^* + \bar{x}_2^*)\right]' S_{11.2}^{-1}(\bar{x}_1^* - \bar{x}_2^*)$$

[2] Fujikoshi and Kanazawa によって与えられた最尤判別統計量

$$Z^* = \frac{N_1}{N_1+1+l_1/k_1}(x^*-\bar{x}_1^*)'S_{11\cdot 2}^{-1}(x^*-\bar{x}_1^*)$$
$$- \frac{N_2}{N_2+1+l_2/k_2}(x^*-\bar{x}_2^*)'S_{11\cdot 2}^{-1}(x^*-\bar{x}_2^*)$$

[3] 同じく Z^* において $l_1=l_2=0$ とした

$$R^* = \frac{N_1}{N_1+1}(x^*-\bar{x}_1^*)'S_{11\cdot 2}^{-1}(x^*-\bar{x}_1^*)$$
$$- \frac{N_2}{N_2+1}(x^*-\bar{x}_2^*)'S_{11\cdot 2}^{-1}(x^*-\bar{x}_2^*)$$

ただし

$x^* = x - S_{12}S_{22}^{-1}y$, $\bar{x}_i^* = \bar{x}_i - S_{12}S_{22}^{-1}\bar{y}_i$ $i=1,2$.

$\bar{x}_i$, $\bar{y}_i$: 標本平均 $S=\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{bmatrix}$ は Σ の最良不偏推定量.

$S_{11\cdot 2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$

$l_i = (y-\bar{y}_i)'S_{22}^{-1}(y-\bar{y}_i)$, $k_i = n/N_i$, $n = N_1+N_2-2$

である.

上の3つの統計量は、次の様な仮説検定問題を考えたときの検定統計量である. 即ち.

H_1 : $\begin{bmatrix} x \\ y\end{bmatrix}, \begin{bmatrix} x_{11} \\ y_{11}\end{bmatrix}, \cdots \begin{bmatrix} x_{1N_1} \\ y_{1N_1}\end{bmatrix}$ は Π_1 に属し.
$\begin{bmatrix} x_{21} \\ y_{21}\end{bmatrix}, \cdots, \begin{bmatrix} x_{2N_2} \\ y_{2N_2}\end{bmatrix}$ は Π_2 に属する。

H_2 : $\begin{bmatrix} x_{11} \\ y_{11}\end{bmatrix}, \cdots, \begin{bmatrix} x_{1N_1} \\ y_{1N_1}\end{bmatrix}$ は Π_1 に属し.
$\begin{bmatrix} x \\ y\end{bmatrix}, \begin{bmatrix} x_{21} \\ y_{21}\end{bmatrix}, \cdots, \begin{bmatrix} x_{2N_2} \\ y_{2N_2}\end{bmatrix}$ に属する.

H_i $(i=1,2)$ の下での尤度関数を L_i とするとき.

[1]. $\mu_1, \mu_2, \eta, \Sigma$ を既知として、L_1, L_2 に μ_1, μ_2, η, Σ の最尤推定量を代入したものが W^* である。

[2]. $\max\limits_{\mu_1,\mu_2,\eta,\Sigma} L_1 \Big/ \max\limits_{\mu_1,\mu_2,\eta,\Sigma} L_2$ が Z^* である。

[3] Σ を既知として、$\max\limits_{\mu_1,\mu_2,\eta} L_1 / \max\limits_{\mu_1,\mu_2,\eta} L_2$ を求め、それに Σ の最尤推定量を代入したものが、R^* になっている。

判別方式は

$W^* > c$ 又は $W^* \leqq c$

$Z^* < k$ 又は $Z^* \geqq k$

$R^* < l$ 又は $R^* \geqq l$

に従って、$\begin{bmatrix} x \\ y \end{bmatrix}$ は、π_1 又は π_2 に属すると判別される。分離点 c, k, l は、0 又は 定数、又は $\bar{x}_i, \bar{y}_i, S$ の関数である。又、最尤判別方式は

$Z^* < 0$ 又は $Z^* \geqq 0$

に従って、$\begin{bmatrix} x \\ y \end{bmatrix}$ は π_1 又は π_2 に属すると判別する方式である。

Covariate 変量がない場合、即ち $q=0$ の場合は、R^* と Z^* は同一の統計量になり、W^* は Anderson の W に、R^*, Z^* は、John-Kudo の Z に対応している。

k_1, k_2 が、それぞれ定数に収束するように N_1, N_2 が大きくなるときの、W^*, Z^*, R^* の極限分布は、

$\begin{bmatrix} x \\ y \end{bmatrix}$ が π_1 に属しているなら、$W^* \sim N[\frac{1}{2}\Delta^2, \Delta^2]$

$Z^*, R^* \sim N[-\Delta^2, 4\Delta^2]$

$\begin{bmatrix} x \\ y \end{bmatrix}$ が π_2 に属しているなら、$W^* \sim N[-\frac{1}{2}\Delta^2, \Delta^2]$

$Z^*, R^* \sim N[\Delta^2, 4\Delta^2]$

に従う。

本稿では、W^*, R^*, Z^* の分布の漸近展開を求め、それらの漸近展開式から得られる、ある種の誤判別の確率の大小を比較することによって、3つの統計量の"よさ"といったものを比較する。

§1. 判別統計量の分布の漸近展開式

[1]. W^*, R^*, Z^* の分布の漸近展開

$$
\begin{aligned}
&\Pr\left\{(W^* - \tfrac{1}{2}\Delta^2)\Delta^{-1} \leqq u \,\middle|\, \begin{bmatrix} x \\ y \end{bmatrix} \in \pi_1\right\} \\
&= \Phi(u) + \phi(u)\Big[\frac{1}{N_1}\Big\{\frac{p}{2\Delta}H_1(u) + \frac{p}{2\Delta^2}H_2(u) + \frac{1}{2\Delta^2}H_4(u)\Big\} \\
&\quad + \frac{1}{N_2}\Big\{-\frac{p}{2\Delta}H_1(u) + \Big(\frac{p}{2\Delta^2} + \frac{1}{2}\Big)H_2(u) - \frac{1}{\Delta}H_3(u) + \frac{1}{2\Delta^2}H_4(u)\Big\} \\
&\quad + \frac{1}{n}\Big\{-\frac{1}{2}(p+q+1)\Delta H_1(u) + \Big(\frac{1}{4}\Delta^2 + \frac{3}{2}(p+q+1)\Big)H_2(u) - \Delta H_3(u) + H_4(u)\Big\}\Big] \\
&\quad + O_2 \qquad : \text{Memon and Okamoto (AMS 1970)}
\end{aligned}
$$

$$\Pr\{(R^*+\Delta^2)(2\Delta)^{-1}\leqq u \mid [\begin{smallmatrix}x\\y\end{smallmatrix}]\in\Pi_1\}$$
$$=\Phi(u)+\phi(u)\Big[\frac{1}{N_1}\Big\{\frac{p}{2\Delta^2}H_2(u)+\frac{1}{2\Delta}H_3(u)+\frac{1}{2\Delta^2}H_4(u)\Big\}$$
$$+\frac{1}{N_2}\Big\{-\frac{\Delta}{2}H_1(u)+\Big(\frac{p}{2\Delta^2}-\frac{1}{2}\Big)H_2(u)+\frac{1}{2\Delta}H_3(u)+\frac{1}{2\Delta^2}H_4(u)\Big\}$$
$$+\frac{1}{m}\Big\{\frac{1}{2}(p+q+1)\Delta H_1(u)+\Big(\frac{3}{2}(p+q+1)+\frac{\Delta^2}{4}\Big)H_2(u)+\Delta H_3(u)$$
$$+H_4(u)\Big\}\Big]+O_2 \qquad : \text{Fujikoshi and Kanazawa}$$

$$\Pr\{(Z^*+\Delta^2)(2\Delta)^{-1}\leqq u \mid [\begin{smallmatrix}x\\y\end{smallmatrix}]\in\Pi_1\}$$
$$=\Phi(u)+\phi(u)\Big[\frac{1}{N_1}\Big\{\frac{p}{2\Delta^2}H_2(u)+\frac{1}{2\Delta}H_3(u)+\frac{1}{2\Delta^2}H_4(u)\Big\}$$
$$+\frac{1}{N_2}\Big\{-\frac{\Delta}{2}H_1(u)+\Big(\frac{p}{2\Delta^2}-\frac{1}{2}\Big)H_2(u)+\frac{1}{2\Delta}H_3(u)+\frac{1}{2\Delta^2}H_4(u)\Big\}$$
$$+\frac{1}{m}\Big\{\frac{1}{2}(p+1)\Delta H_1(u)+\Big(\frac{1}{2}(3p+q+3)+\frac{\Delta^2}{4}\Big)H_2(u)+\Delta H_3(u)$$
$$+H_4(u)\Big\}\Big]+O_2 \qquad : \text{Fujikoshi and Kanazawa}$$

$[\begin{smallmatrix}x\\y\end{smallmatrix}]\in\Pi_2$ の場合は、上の各式の N_1 と N_2 を入れ換えれば得られる。又、

$$H_1(u)=1,\quad H_2(u)=-u,\quad H_3(u)=u^2-1,\quad H_4(u)=-u^3+3u$$

である。

[2] Studentized W^*, R^*, Z^* の分布の漸近展開

$D^2=(\bar{x}_1^*-\bar{x}_2^*)'S_{11\cdot2}^{-1}(\bar{x}_1^*-\bar{x}_2^*)$ は 0 でないと仮定する。

$$\Pr\{(W^* - \tfrac{1}{2}D^2)D^{-1} \leqq u \mid [{x \atop y}] \in \Pi_1\}$$
$$= \Phi(u) + \phi(u)\Big[\frac{1}{N_1}\Big\{\frac{1}{\Delta}(p-1)H_1(u) + \frac{1}{2}H_2(u)\Big\}$$
$$+ \frac{1}{n}\Big\{(p+q)H_2(u) + \frac{1}{4}H_4(u)\Big\}\Big] + O_2$$

: Kanazawa and Fujikoshi

$$\Pr\{(R^* + D^2)(2D)^{-1} \leqq u \mid [{x \atop y}] \in \Pi_1\}$$
$$= \Phi(u) + \phi(u)\Big[\frac{1}{N_1}\Big\{-\frac{1}{2\Delta}(p-2)H_1(u) + \frac{1}{2}H_2(u) + \frac{1}{2\Delta}H_3(u)\Big\}$$
$$+ \frac{1}{N_2}\Big\{-\Big(\frac{p}{2\Delta} + \frac{\Delta}{2}\Big)H_1(u) - H_2(u) - \frac{1}{2\Delta}H_3(u)\Big\}$$
$$+ \frac{1}{n}\Big\{(p+q)H_2(u) + \frac{1}{4}H_4(u)\Big\}\Big] + O_2$$

: Fujikoshi and Kanazawa

$$\Pr\{(Z^* + D^2)(2D)^{-1} \leqq u \mid [{x \atop y}] \in \Pi_1\}$$
$$= \Phi(u) + \phi(u)\Big[\frac{1}{N_1}\Big\{-\frac{1}{2\Delta}(p-2)H_1(u) + \frac{1}{2}H_2(u) + \frac{1}{2\Delta}H_3(u)\Big\}$$
$$+ \frac{1}{N_2}\Big\{-\Big(\frac{p}{2\Delta} + \frac{\Delta}{2}\Big)H_1(u) - H_2(u) - \frac{1}{2\Delta}H_3(u)\Big\}$$
$$+ \frac{1}{n}\Big\{-\frac{q\Delta}{2}H_1(u) + pH_2(u) + \frac{1}{4}H_4(u)\Big\}\Big] + O_2$$

: Fujikoshi and Kanazawa

$[{x \atop y}] \in \Pi_2$ の場合には、上の各式の N_1 と N_2 を入れ換えればよい。

§2. 第一種の誤判別の確率を一定にしたときの第二種の誤判別の確率

ここでは，Studentized W^*, R^*, Z^* の分布の漸近展開を用いて，$\begin{bmatrix} x \\ y \end{bmatrix}$ が π_1 に属しているのに，π_2 に属していると誤って判別する確率が与えられた値：α になるように，それぞれの分離点を決めたとき，$\begin{bmatrix} x \\ y \end{bmatrix}$ が π_2 に属しているのに，誤って π_1 に属していると判別される確率を求める。

$$\Pr\left\{ W^* \leqq c_\alpha \,\middle|\, \begin{bmatrix} x \\ y \end{bmatrix} \in \pi_1 \right\} = \alpha, \quad \Pr\left\{ R^* \geqq e_\alpha \,\middle|\, \begin{bmatrix} x \\ y \end{bmatrix} \in \pi_1 \right\} = \alpha,$$

$$\Pr\left\{ Z^* \geqq k_\alpha \,\middle|\, \begin{bmatrix} x \\ y \end{bmatrix} \in \pi_1 \right\} = \alpha$$

となる分離点，c_α, e_α, k_α は，次の漸近展開式によって与えられる。

$$c_\alpha = \xi_W D + \frac{1}{2} D^2, \quad e_\alpha = 2\xi_R D - D^2, \quad k_\alpha = 2\xi_Z D - D^2$$

ただし

$$\xi_W = u_0 - \frac{1}{2N_1\Delta}\{2(p-1)\Delta H_2(u_0)\} - \frac{1}{4n}\{4(p+q)H_2(u_0) + H_4(u_0)\} + O_2$$

$$\xi_R = -u_0 + \frac{1}{2N_1\Delta}\{(p-2)H_1(u_0) + \Delta H_2(u_0) - H_3(u_0)\} + \frac{1}{2N_2\Delta}\{(p+\Delta^2)H_1(u_0) - 2\Delta H_2(u_0) + H_3(u_0)\} + \frac{1}{4n}\{4(p+q)H_2(u_0) + H_4(u_0)\} + O_2$$

$$\xi_Z = -u_0 + \frac{1}{2N_1\Delta}\{(p-2)H_1(u_0) + \Delta H_2(u_0) - H_3(u_0)\}$$
$$+ \frac{1}{2N_2\Delta}\{(p+\Delta^2)H_1(u_0) - 2\Delta H_2(u_0) + H_3(u_0)\}$$
$$+ \frac{1}{4n}\{2q\Delta H_1(u_0) + 4pH_2(u_0) + H_4(u_0)\} + O_2$$

ただし，u_0 は $N[0,1]$ の 100α %点である．これらの分離点を用いたとき，

$$Pr\{W^* > c_\alpha \mid [\begin{smallmatrix}x\\y\end{smallmatrix}] \in \pi_2\}$$
$$= 1 - \Phi(u_0+\Delta) - \phi(u_0+\Delta)\Big[\frac{1}{2N_1\Delta}\{\Delta^2 - (p-1)\} - \frac{1}{2N_2\Delta}\{4(u_0+\Delta)\Delta - 3(p-1)\} - \frac{\Delta}{4n}\{u_0^2 + 2(p+q-1)\}\Big] + O_2$$

$$Pr\{R^* < e_\alpha \mid [\begin{smallmatrix}x\\y\end{smallmatrix}] \in \pi_2\}$$
$$= 1 - \Phi(u_0+\Delta) - \phi(u_0+\Delta)\Big[\frac{1}{2N_1\Delta}\{2u_0^2 + 4\Delta u_0 + 3\Delta^2 + (p-1)\}$$
$$- \frac{1}{2N_2\Delta}\{2(u_0+\Delta)(u_0+3\Delta) - (p-1)\} - \frac{\Delta}{4n}\{u_0^2 + 2(p+q-1)\}\Big]$$
$$+ O_2$$

$$Pr\{Z^* < k_\alpha \mid [\begin{smallmatrix}x\\y\end{smallmatrix}] \in \pi_2\}$$
$$= 1 - \Phi(u_0+\Delta) - \phi(u_0+\Delta)\Big[\frac{1}{2N_1\Delta}\{2u_0^2 + 4\Delta u_0 + 3\Delta^2 - (p-1)\}$$
$$- \frac{1}{2N_2\Delta}\{2(u_0+\Delta)(u_0+3\Delta) - (p-1)\} - \frac{\Delta}{4n}\{u_0^2 + 2(p-q-1)\}\Big]$$
$$+ O_2$$

これより、すぐに言えることは、共変量の次元qが大きければ、R^*とZ^*を比べると、Z^*の方が、第二種の誤確率が小さくなるということである。又、この三つの判別誤確率の中、一番小さい確率を与えるのは、次の表の統計量である。三つの統計量の大きさの表にすると、もっと細かく場合分けしなければならないので、次表にとどめる。

<table>
<tr><th rowspan="2">Δ \ k_2</th><th rowspan="2">$k_2 \leqq k_1$</th><th rowspan="2">$k_1 < k_2 < 2k_1+q$</th><th rowspan="2">$2k_1+q = k_2$</th><th colspan="2">$2k_1+q < k_2$</th></tr>
<tr><th>$u_0 < u^*$</th><th>$u^* \leqq u_0$</th></tr>
<tr><td>0</td><td rowspan="4">Z^*</td><td rowspan="2">W^*</td><td>W^*</td><td rowspan="2">W^*</td><td rowspan="4">W^*</td></tr>
<tr><td></td><td rowspan="3">Δ_0
Z^*</td></tr>
<tr><td></td><td rowspan="2">Δ_2
Z^*</td><td>Δ_2
Z^*</td></tr>
<tr><td>大</td><td>Δ_1
W^*</td></tr>
</table>

ただし、$u^* = -\{(k_2-2k_1-q)(p-1)/(k_1+q)\}^{1/2}$, $\Delta_0 = -(u_0^2+p-1)/2u_0$,

$\Delta_1 = \{-(k_2-k_1)u_0+h\}/(k_2-2k_1-q)$,

$\Delta_2 = \{-(k_2-k_1)u_0-h\}/(k_2-2k_1-q)$,

$h^2 = (k_2-k_1)\{(k_1+q)u_0^2-(k_2-2k_1-q)(p-1)\}$ である。

判別問題においては、Δが小さいときの方が判別しにくいわけであるが、Δが小さいとき、片方の誤確率を一定にするという条件の下では、線形判別統計量の方が、より好ましいという結論になる。なお、この表は、W^*, R^*, Z^*の漸近展

開式 §1. [1] から計算しても、同じ結果が得られる。又、共変量を無視して、この判別問題を、普通の p-次元判別問題として、取り扱うと、

$$\Delta_0^2 = (\mu_1 - \mu_2)' \Sigma_{11}^{-1} (\mu_1 - \mu_2) < \Delta^2$$

であるから、第二種の誤判別の確率は、一般に大きくなる。

参考文献

[1] Cochran, W.G. and Bliss, C.I. (1948). Discriminant functions with covariance. Ann. Math. Stat. 19 151-176.

[2] Cochran, W.G. (1964). Comparison of two methods of handling covariates in discriminatory analysis. Ann. Inst. Statist. Math. 16 43-53.

[3] Memon, A.Z. and Okamoto, M. (1970). The classification statistic W* in covariate discriminant analysis. Ann. Math. Stat. 41 1491-1499.

[4] Kanazawa, M. and Fujikoshi, Y. (1975). An asymptotic expansion of the distribution of the studentized classification W* and its applications.

[5] Fujikoshi, Y. and Kanazawa, M. (1974). The maximum likelihood classification statistic in covariate discriminate analysis and its applications.